International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 5, Issue 3, Jun 2015, 23-30
TJPRC Pvt. Ltd.

AUTOMATIC EXTRACTION OF TOP-K LISTS FROM THE WEB PAGES

ASHISH DWIVEDI

(M.E.(Computer)) Lokmanya Tilak College of Engineering & Technology, LTCOE, NAVI Mumbai, India

ABSTRACT

The World Wide Web is currently the largest source of information. The information on the Web is structured (such as table) and unstructured or semi-structured form. Unfortunately, in most cases, context is expressed in structured text that machines can not interpret. Therefore this project is concerned with information extraction from top-k Web pages (rich and valuable source of information), which are web pages that describe top k instances of a topic which is of general interest.

Examples include “20 Most Influential Scientists Alive Today”, “the 10 hits of 2010 you don’t want to miss” etc. Compared to other structured information on the web (including web tables), information in top-k lists is larger and richer, of higher quality, generally more interesting. Therefore top-k list are highly valuable. For application such as search or fact answering, it can help enrich open-domain knowledge bases for their support. The paper concentrate on an efficient method, called Tag Path Clustering that extracts top-k list from web pages with high performance.

KEYWORDS: Web Lists, Web Information Extraction, Top-k Lists, Lists Extraction, Web Mining

INTRODUCTION

The World Wide Web is currently the largest source of information. However, most information on the web is unstructured text in natural languages, and extracting knowledge from natural language text is very difficult. Still, some information on the web exists in structured or semi-structured forms, for example, as lists or web tables coded with specific tags such as <ul>, <li>, and <table> on html pages. As a result, a lot of recent work has focused on acquiring knowledge from structured information on the web, in particular, from web tables. In this paper, instead of focusing on structured data (such as tables) and ignoring context, we focus on context that we can understand, and then we use the context to interpret less structured or almost free-text information, and guide their extraction. Specifically, we focus on a rich and valuable source of information on the web, which we call top-k web pages. A top-k web page describes k items of a particular interest. Some typical titles are:

- 20 Most Influential Scientists Alive Today
- Twelve Most Interesting Children’s Books in USA
- 10 Hollywood Classics You Shouldn’t Miss

The title of a top-k page contains at least three pieces of important information: i) A number k, for example, 20, Twelve which indicates how many items are described in the page; ii) A topic or concept the items belong to, for example, Scientists, Children’s Books iii) A ranking criterion, for example, Influential, Interesting. Some top-k titles contain two

optional pieces of information: time and location. For example, 2011 and USA in the above example. In this paper, we present an efficient method that extracts top-k lists from web pages with high performance.

EXISTING SYSTEM

The World Wide Web is currently the largest source of information. A lot of recent work has focused on acquiring knowledge from structured information on the web, in particular, from web tables. However, it is questionable how much valuable knowledge we can extract from lists and web tables. It is true that the total number of web tables is huge in the entire corpus, but only a very small percentage of them contain useful information. An even smaller percentage of them contain information interpretable without context. According to, among the 1.1% of all web tables are relational, a lot of them are meaningless without context. For example, suppose we extracted a table that contains 5 rows and 2 columns, with the 2 columns labeled “Companies” and “Revenue” respectively. It is still unclear why these 5 companies are grouped together (e.g., are they the most profitable, most innovative, or most employee friendly companies of a particular industry in a particular region), and how should we interpret their revenues (e.g., in which year or even in what currency). In other words, we do not know in what circumstances people will find the extracted information interesting or useful. It is clear that understanding the context is extremely important in information extraction. Unfortunately, in most cases, context is expressed in unstructured text that machines cannot interpret.

PROPOSED SYSTEM

In this paper, instead of focusing on structured data (such as tables) and ignoring context, we focus on context that we can understand, and then we use the context to interpret less structured or almost free-text information, and guide their extraction. Specifically, we focus on a rich and valuable source of information on the web, which we call top-k web pages. A top-k web page describes k items of a particular interest. In most cases, the description is in natural language text which is not directly machine interpretable, although the description has the same format or style for different items. But most importantly, the title of a top-k page often clearly discloses the context, which makes the page interpretable and extractable.

MODULES DESCRIPTION

- Title Classifier
- Creating a training dataset
- Extracting features
- Candidate Picker
- Top-K Ranker.

Title Classifier

The title of a web page (string enclosed in <title> tag) helps us identify a top-k page. There are several reasons for us to utilize the page title to recognize a top-k page. First, for most cases, page titles serve to introduce the topic of the main body. Second, while the page body may have varied and complex formats, top-k page titles have relatively similar structure. Also, title analysis is lightweight and efficient. If title analysis indicates that a page is not a top-k page, we chose to skip this page. This is important if the system has to scale to billions of web pages.

Figure 1: A Sample og Top k Title

Creating a Training Dataset

Creating a large, high quality training dataset is costly. The challenge mainly lies in collecting positive cases, as top-k pages are sparse on the web (approx. 1.4‰ of total web pages). Filtering out pages without a number in the title narrows our candidates down, but the number of candidates is still massive. In our approach, we first parse the titles to add POS tags, and then we adopt the following simple rules to identify or create positive training samples.

• "top CD"

If a title contains the word "top" followed by a number, it is likely to be top-k title. For example, "top 10 NBA players who could be successful general managers".

• "top CD" without "top"

A title which satisfies the "top CD" rule is still a top-k title with the word "top" removed.

• "CD JJS": "JJS"

X stands for superlative adjectives. If a title contains a number followed by a superlative adjective, it is likely to be a top-k title. For example, "20 tallest buildings in China".

• "CD RBS JJ"

"RBS" and "JJ" stand for superlative adverbs and adjectives, respectively. If a title contains a number, followed by a superlative adverb, and followed by an adjective, it is likely to be a top-k title. For example, "5 most expensive watches in the world".

Table 1: Feature Extraction from a Window of Size 9. (Vacancies are Filled with the Null Token)

word	lemma	POS	concept	tag
.net	net	JJ	1	FALSE
awards	award	NNS	1	FALSE
2011	2011	CD	0	FALSE
top	top	JJ	1	FALSE
10	10	CD	0	TRUE
podcasts	podcast	NNS	1	FALSE
NULL	NULL	NULL	NULL	FALSE
NULL	NULL	NULL	NULL	FALSE
NULL	NULL	NULL	NULL	FALSE

Extracting Features

We now discuss how we extract features from a title. As we see in Figure 3, a title may contain multiple segments, which are separated by separators like "-" or "|". Among these segments, only the main segment (e.g, Segment 1 in Figure 3) gives us the topic of the page, while other segments show additional information such as the name of the site, which is not of interest. We therefore split the title and retain only segments that contain a number. Instead of extracting features from a title as a whole, we focus on a fixed-size window centered around the number k in the title. We argue that the number k serves as an anchor to a phrase that represents a top-k concept or topic. For a window of large enough size n,

the n-gram is sufficient to make a correct judgement. With this observation, we transform the original task into the task of recognizing the number k with a proper context, which is much easier and more suitable for CRF learning.

Candidate Picker

This step extracts one or more list structures which appear to be top-k lists from a given page. A top-k candidate should first and for most be a list of k items, Visually, it should be rendered as k vertically or horizontally aligned regular patterns. While structurally, it is presented as a list of HTML nodes with identical tag path. A tag path is the path from the root node to a certain tag node, which can be presented as a sequence of tag names. Figure 6 shows the relation between list nodes and tag paths. Based on these observations, the system employs two basic rules for selecting candidate lists:

- **K Items:** A candidate list must contain exactly k items.
- **Identical Tag Path:** The tag path of each item node in a candidate list must be the same.

1. **Index:** There exists an integer number in front of every list item, serving as a rank or index: e.g., “1.”, “2.”, “3.”, etc. Moreover, the numbers are in sequence and within the range of [1, k]. List Nodes and Their Tag Paths
2. **Highlighting Tag:** The tag path of the candidate list contains at least one tag among <b>,,<h1-h6> for highlighting purposes.
3. **Table:** The candidate list is shown in a table format.

Top-K Ranker

Top-K Ranker ranks the candidate set and picks the top ranked list as the top-k list by a scoring function which is a weighted sum of two feature scores below:

- **P -Score:** P -Score measures the correlation between the list and title. We extract a set of concepts from the title, and one of them is the central concept of the top-k list. Our key idea is that one or more items from the main list should be instances of that central concept from the title.
- **V -Score:** V -Score calculates the visual area occupied by a list, since the main list of the page tends to be larger and more prominent than other minor lists. The V –Score of a list is the sum of the visual area of each node and is computed.

RESULTS

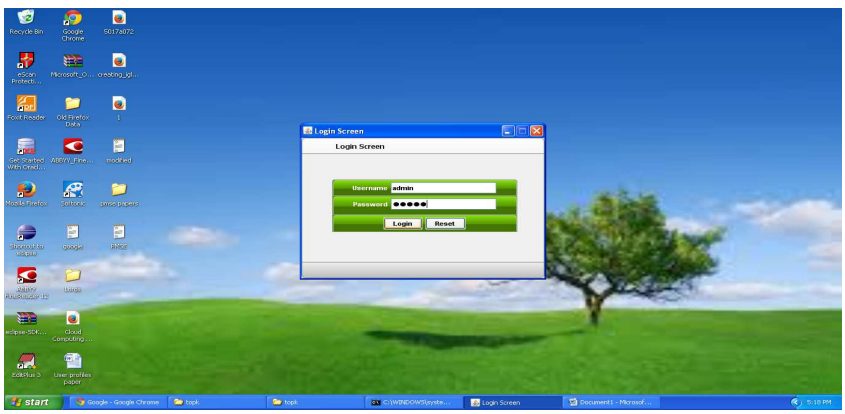

Figure 2: Login screen After Successfully Login

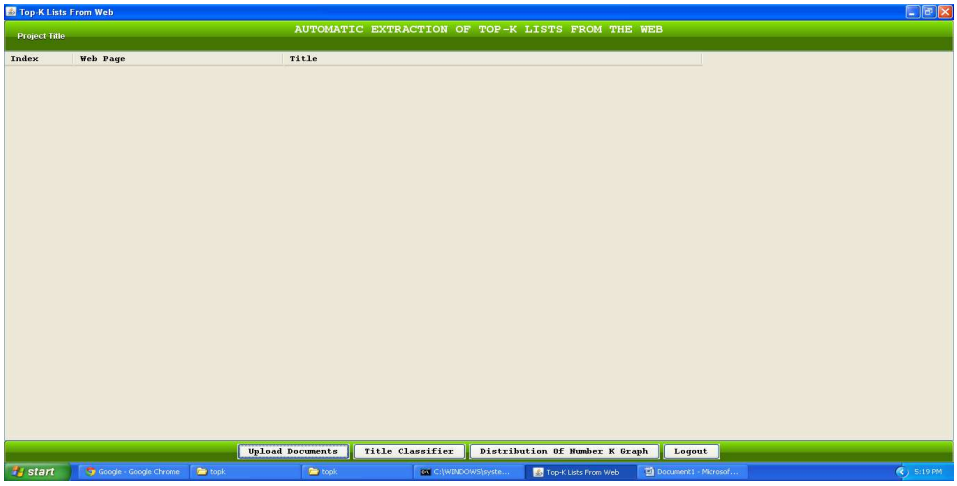

Figure 3: Uploading a Dataset to Extract the Top-K Lists,

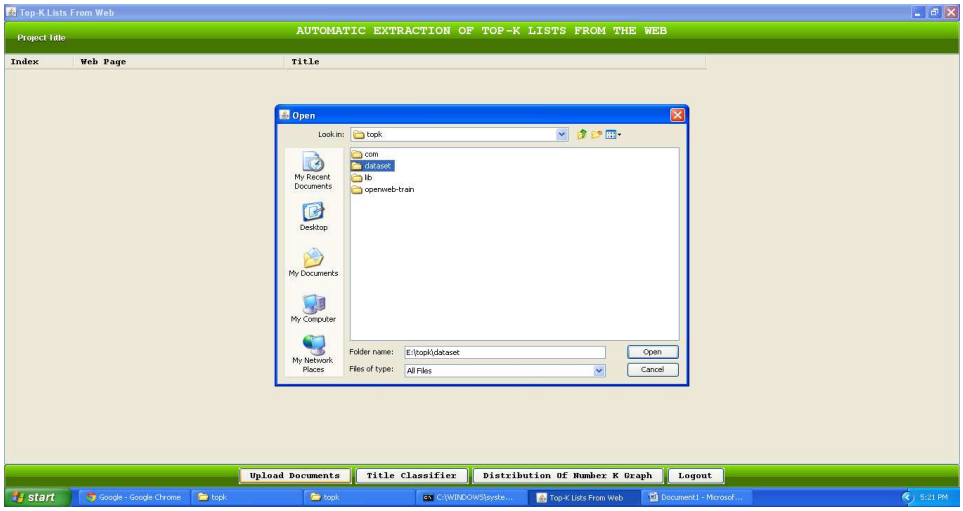

Figure 4: Uploading a Dataset

Index	Web Page	Title
0	0.html	Thursday Thirteen – Fairy Tale Lessons | Lives Less Ordinary
1	1.html	Which continent has the highest life expectancy
2	2.html	Bratz - Wikipedia, the free encyclopedia
3	3.html	100 novels everyone should read - Telegraph
4	4.html	List of alcoholic beverages - Wikipedia, the free encyclopedia
5	5.html	List of home computers - Wikipedia, the free encyclopedia
6	6.html	IMDb: Upcoming movies: 2013, 2014, 2015 and beyond - a list by felcabo
7	7.html	CNN.com - September 11 Memorial
8	8.html	List of Los Angeles Unified School District schools - Wikipedia, the fr...
9	9.html	List of active volcanos - Simple English Wikipedia, the free encyclopedia
10	10.html	General Surgeon Doctors in Kolkata, Doctors General Surgeon, India | Ju...
11	11.html	Today (U.S. TV program) - Wikipedia, the free encyclopedia
12	12.html	Category:Amphetamines - Wikipedia, the free encyclopedia
13	13.html	Tier 4 Sponsors Register | List of sponsors (UK Universities, Colleges...
14	14.html	21 Things Republicans Have Demanded To Avoid A Crisis
15	15.html	List of semiconductor materials - Wikipedia, the free encyclopedia
16	16.html	Environmental Threats Information, Environmental Threats Facts, Environ...
17	17.html	Disney Princess - Wikipedia, the free encyclopedia
18	18.html	IMDb: Funniest Comedy Movies Since 2000 [Top 70] - a list by Dan T
19	19.html	List of Harry Potter Movies in Order | Harry Potter Spells

Figure 5: After Successful Loading of Dataset

Now click on Title Classifier to extract the features (The Features were extracted according to, from title t we can extract a 5-tuple (k, c, m, t, l) where k is a natural number, c is a noun-phrase concept defined in a knowledge base such as the one described, m is a ranking criterion, t is temporal information, l is location information. Note that k and c are mandatory, while m, t, and l are optional.

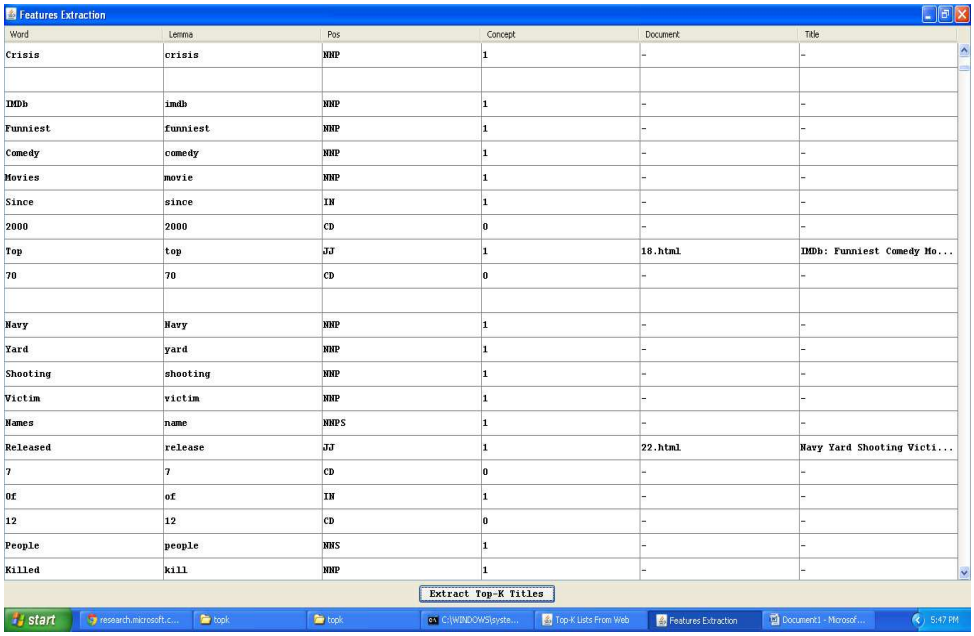

Figure 6: Here in This Top –k LIST will be Extracted Based on POS Values where CD Followed by JJ, CD Followed by JJS Those will be Treated as Positive Title and will be Added in the List of Top-K List

Figure 7: The Titles are classified according to Title Classifier, and classified List will be Treated as Top-K List

From this list you will able to get the candidates available in that link, for that Click on

Candidate Picker. Here it is showing the Top 100 players in MLB according to the title.

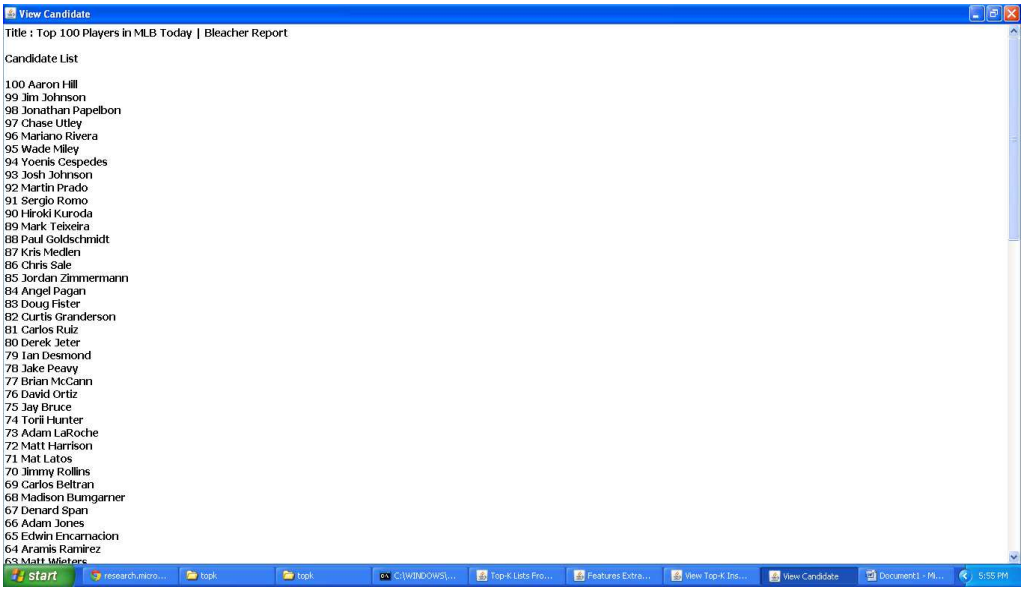

Figure 8: The Candidate List Output Showing the Top 100 Players in MLB According to The title

CONCLUSIONS

This paper presents a novel and interesting problem of extracting top-k lists from the web. Compared to other structured data, top-k lists are clearer, easier to understand and more interesting for human consumption, and therefore are an important source for data mining and knowledge discovery. This project also attempts to study & compare different approaches for extracting top-k data lists from the web pages. It also studies limitation of each approach and how it is bridged in other subsequent approaches. We demonstrate an algorithm that automatically extracts over 1.7 million such lists from the a web snapshot and also discovers the structure of each list.

REFERENCES

1. ".net awards 2011: top 10 podcasts," http://goo.gl/9D8vj.
2. "Google sets," http://labs.google.com/sets.
3. M. J. Cafarella, E. Wu, A. Halevy, Y. Zhang, and D. Z. Wang, "Webtables: Exploring the power of tables on the web," in VLDB, 2008.
4. B. Liu, R. L. Grossman, and Y. Zhai, "Mining data records in web pages," in KDD, 2003, pp. 601–606.
5. G. Miao, J. Tatemura, W.-P. Hsiung, A. Sawires, and L. E. Moser, "Extracting data records from the web using tag path clustering," in WWW, 2009, pp. 981–990.
6. W. Gatterbauer, P. Bohunsky, M. Herzog, B. Kr ¨ upl, and B. Pollak, "Towards domain-independent information extraction from web tables," in WWW. ACM Press, 2007, pp. 71–80.
7. F. Fumarola, T. Weninger, R. Barber, D. Malerba, and J. Han, "Extracting general lists from web documents: A hybrid approach," in IEA/AIE (1), 2011, pp. 285–294.
8. W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in SIGMOD, 2012.